Tail universalities in rank distributions as an algebraic problem: the beta-like function

G.G. Naumis

Departamento de Fisica-Quimica, Instituto de Fisica. Universidad Nacional Autónoma de México. Apdo. Postal 20-364, 01000, México D.F., Mexico.

G. Cocho

Departamento de Sistemas Complejos, Instituto de Fisica. Universidad Nacional Autónoma de México. Apdo. Postal 20-364, 01000, México D.F., Mexico.

Abstract

Although power laws of the Zipf type have been used by many workers to fit rank distributions in different fields like in economy, geophysics, genetics, soft-matter, networks etc., these fits usually fail at the tails. Some distributions have been proposed to solve the problem, but unfortunately they do not fit at the same time both ending tails. We show that many different data in rank laws, like in granular materials, codons, author impact in scientific journal, etc. are very well fitted by a beta-like function. Then we propose that such universality is due to the fact that a system made from many subsystems or choices, imply stretched exponential frequency-rank functions which qualitatively and quantitatively can be fitted with the proposed beta-like function distribution in the limit of many random variables. We prove this by transforming the problem into an algebraic one: finding the rank of successive products of a given set of numbers.

Key words: Ranking distributions, Power law distribution, Zipf law, Multiplicative processes
PACS: : 89.75.Fb, 87.10.+e, 89.75.Da, 89.65.Gh, 89.65.-s, 87.23.Cc

1 Introduction

Both natural language texts and coding DNA sequences present power laws in the observed frequency of a word as a function of its rank (r), where the

Email address: naumis@fisica.unam.mx (G.G. Naumis).

rank is just the ordinal position of a word if all words are ordered according to their decreasing frequency. Usually, the most frequent word has rank 1, the next most frequent rank 2 and so on. This power law behavior of the ranking is known as the Zipf law [1], and it is very common in physics, biology, geography, economics, linguistics, etc. [1]. In physics one can cite the rank distribution of stick-slip events in sheared granular media [2], earthquakes [2], radionuclides half-life time and nuclides mass number [3]. Many complex systems share as well the same phenomenology, as happens in networks [4], biological clocks [5] and metabolic networks [6]. Zipf discovered his rank law by analyzing manually the frequencies of 29,899 different words types in the novel "Ulysses" by James Joyce, but when a larger set of words is considered, a deviation from a power law is observed for larger ranks [7]. A similar behavior is found in coding genetic sequences. Deviations from the Zipf law are also found in the tails ranking of many physical systems [8]. In fact, is clear that one should expect a different behavior at the tails, since finite size effects should be present and the power law must be "stopped" at a certain region. In spite of this, many workers just ignore the tail effects by fitting the data in a restricted range, or they proceed in a very questionable way by fitting all the data with a power law. Others have fitted sets of data in nature and in economy with stretched exponentials [8] and log-normal distributions [9]. The problem with the previous expressions is that they do not fit the data at *both ending tails*, where different kinds of processes are set in once a crossover region is reached. Such crossovers are due to finite size effects, in which different mechanisms are set in when certain big and small scales are reached. This leads to the idea of using multiscaling physical modelling to understand such features. Maybe the best example of the previous situation occurs in turbulence, where Kolmogorov´s power law is observed only in the inertial regimen [10][11]. In one tail (small length scales) energy dissipation plays the main role, while energy injection dominates at big scales [10][11]. For each of these limits, the scaling behavior is different [12][13]. One can conjecture that similar ideas are behind many other complex physical systems, since we report that many rank laws are extremely well parametrized, outperforming many other rank-order models, with a two exponent beta function-like formula with parameters $\{a, b\}$,

$$f(r) = K\frac{(R-r+1)^b}{r^a}, \tag{1}$$

where a and b are fitted from the data, r is the rank and R is the maximal r. If $f(r)$ is normalized to 1, then $K \equiv 1/\sum_{r=1}^{N}(R-r+1)^b/r^a$. For $R \gg 1$, K can be transformed into an integral that yields $K \approx \Gamma(b-a+2)/\Gamma(1-a)\Gamma(1+b)$. We will show that $f(r)$ is related with a kind of central limit theorem, in which a and b seem to be parameters related with the onset of different mechanisms. Our work is in the same spirit of Moyano *et. al.* [14], who have commented that the rather ubiquitous presence of the Tsallis q-distributions is maybe due to a q-generalized central limit theorem for a class of non independent, correlated, product of probability distributions [15]. The outline of this pa-

Fig. 1. Population ranking of four representative municipalites from Mexico and Spain. The solid lines are the fits obtained from Eq. (1) The inset presents the corresponding values of a and b used in the fits.

per is the following: in section II we present some representative examples of the phenomenology that we have observed. In section III we show how this phenomenology can be studied as a problem of hierarchies in the product of random variables, and then transformed into a related algebraic problem: what is the rank of a set of numbers produced by the iterative product of an initial finite set of numbers. In section IV, we solve the proposed problem, and finally, in section V we give the conclusions of this work.

2 Phenomenology of rank laws and the beta-like function

As starting point, we will provide some representative results of the wide phenomenology found in the tails of rank laws. We start with an example from geography. Fig. 1 shows the population ranking of four representative municipalities in Mexico and Spain in a semilog plot. The corresponding fits using Eq. (1) are given by solid lines. The agreement is excellent, with a correlation coefficient R bigger than 0.98 for all fits. The values of a and b for each fit are shown in the inset of the plot. We have verified that similar good results are obtained for the population of countries and states.

Figure 2 shows the impact factor against the rank of scientific journals, taken from a recent study [16], compared with the fits given by Eq. (1). Again, all the fits are excellent, with correlation coefficients above 0.98.

Similar excellent fitting results are obtained for codon usage in genomes, as shown in Fig. 3, where we plot the logarithm of the frequency of codons (normalized to 1000) as a function of the rank for different representative

Fig. 2. Impact factor as a function of the rank for physics, computer science and agroscience. Fits using the beta-like function are shown as solid lines. Inset: values of a and b.

Fig. 3. Frequency of codons (normalized to 1000) as a function of the rank for the genome of four different species, with their correponding fits shown as solid lines. Inset: values of a a and b used for the fits in the beta-like distribution.

organisms, taken from a genome database [17]. For all the organisms, the resulting correlation parameters are bigger than 0.97.

Now we turn our attention to physics. In Fig. 4 we plot the rank-ordered distribution of stick-slip events in a slowly sheared granular media taken from Ref. [2], fitted using Eq. (1). Although a modified power law was proposed in Ref. [2] to explain the results, the present fit also gives a better correlation coefficient.

Here we presented four examples, but Eq. (1) can be used with excellent results in order to correct the Gutenberg-Ritcher law in earthquakes ranking, Bénard convection cells and in many different fields, like arquitecture, music or roads [18].

Fig. 4. Rank-ordered distribution of stick-slip events in a slowly sheared granular media. Circles are data taken from Ref. [2], and the solid line is a fit using Eq. (1), with $a = 1.08$ and $b = 0.40$

3 Hierarchy in a multiplicative stochastic processes

The previous section leads to the conclusion that both ending tails of the ranking present some degree of universality, and Eq. (1) seems to be an excellent fitting function due to the fact that it gives the right shape of the curve and thus very good correlation coefficients. Also, it is simple and can be reduced to a pure power law by using an appropriate choice of a and b. As the $\{a, b\}$ distributions is indeed ubiquitous, one can try to associate it to some generic mechanism, as happens in the central limit theorem or in the product of correlated probability distributions [14].

In the dynamics of population, scientific journal impact factor, codon usage and stick-slip events, there are many important issues that determine the behavior. In the case of the impact factor we can cite for example the ability to select a good problem for investigation, the gift for writing clear papers, etc. Similar comments would be valid for the dynamics of granular media, as well as in economy, linguistics, genetics, etc. All of the previous systems share a common feature: their complex nature, *i.e.*, they are build from many subsystems or path choices that produce a final result. One can try to model such complexity as follows. Consider a system made from N identical subsystems, where each can have s different states or choices with probability p_j, and $j = 1, ..., s$. When N such subsystems are put together, the state space consists of all s^N possible sequences of length N. If we do not care about the order of the choices or states in the string, there are just $(N+s-1)!/s!(N-1)!$ *different combinations*. For example, if a system is made from $N = 2$ subsystems, where each has two states or choices, say 1 or 0, the possible global states are $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$, while there are only three combina-

tions: $(0,0)$, $(1,1)$ and $(1,0)$, the last one has multiplicity 2. Each combination has a certain probability that we call *reduced probabilities* $x_N(n_1, n_2, ..., n_s)$. The multiplicity of each different state is given by the multinomial coefficient $N!/(n_1!n_2!n_3!...n_s!)$, where n_j is the number of subsystems in the j-esim state. The probability of a global state of the whole system is,

$$P_N(n_1, n_2, ..., n_s) = \frac{N!}{n_1!n_2!n_3!...n_s!} x_N(n_1, n_2, ..., n_s), \tag{2}$$

with $n_1 + n_2 + n_3 + ...n_s = N$. However, we are interested in the rank of the observed different values of the macrostates, not in their distribution of probability. To tackle this problem, we notice that each value $x_N(n_1, n_2, ..., n_s)$ corresponds to a *different macrostate of the system*. In our example, the states $(0,1)$ and $(1,0)$ produce the same global macrostate. These two internal states lead to one global state that has the same characteristics. If one assume that a certain characteristic (X) of a process or object is a function of $n_1, n_2, ..., n_s$, then each value of $X(n_1, n_2, ..., n_s)$ can be mapped to $x_N(n_1, n_2, ..., n_s)$ and $X(n_1, n_2, ..., n_s) = X(x_N(n_1, n_2, ..., n_s))$. From the previous considerations, is clear that any rank hierarchy of $x_N(n_1, n_2, ..., n_s)$ will be inherited to $X(n_1, n_2, ..., n_s)$. Thus, many different rank features of a system are reduced to study the hierarchy present in $x_N(n_1, n_2, ..., n_s)$.

For doing such study, there are two cases. In the first, the subsystems are independent, as in a Bernoulli process,

$$x_N(n_1, n_2, ..., n_s) = p_1^{n_1} p_2^{n_2} p_3^{n_3} ... p_s^{n_s}, \tag{3}$$

and the other is the general case of interacting subsystems, in which the addition of a new subsystem leads to a functional relationship of the type,

$$x_{N+1}(n_1, n_2, ..., n_s) = f(x_N(n_1, n_2, ..., n_s)). \tag{4}$$

In the next section we will consider only the case of independent subsystems, in which no extra information is needed in order to model the correlation in the system. This allows to produce the beta-like function in a simple form.

4 The rank hierarchy as an algebraic problem

For independent subsystems, an inspection of Eq. (3) shows that the rank structure can be reduced to the following algebraic problem. Take s numbers

Fig. 5. Succesive multiplication of three numbers $p_1 = 0.5202, p_2 = 0.3125$, $p_3 = 0.1673$ as a function of the rank (bold solid line), and a fitting using Eq. (1), with $a = 9.36$, $b = 14.53$.

p_1, p_2, ..., p_s at random (normalization can be imposed at the end of the process), labeled in such a way that $p_1 > p_2 > ... > p_s$, and multiply once each number by all the numbers in the set. With these resulting numbers, repeat the process N times to obtain a set of numbers that have the form $p_1^{n_1} p_2^{n_2} p_3^{n_3} ... p_s^{n_s}$, where $n_1 + n_2 + ... + n_s = N$. If the resulting numbers are arranged in decreasing magnitude, we can assign a rank (r) to each one according to its order in the hierarchy. The rank $r = 1$ is assigned to p_1^N, while the lowest rank $r = R$ corresponds to p_s^N. For example, chose at random three numbers p_1, p_2 and p_3 and form all the possible products: $p_1^2, p_1 p_2, p_1 p_3, p_2^2, p_2 p_3, p_3^2$. In Fig. 5, we present a plot of $\log x_N(n_1, n_2, n_3)$ as a function of r for $N = 30$ and $p_1 = 0.5202$, $p_2 = 0.3125$ and $p_3 = 0.1673$. Fig. 5 shows that the resulting ranks are well fitted by the same two parameter beta-like function, with $a = 9.36 \pm 0.2$ and $b = 10.52 \pm 0.2$, with a correlation coefficient of 0.972. The message from this numerical experiment is simple: if this product is seen as a multiplicative process where each number is the probability of making a certain choice or state in a process, then each possible result has a well determined hierarchy.

The task that remains is how to calculate $x_N(n_1, n_2, ..., n_s)$ in terms of the rank. The problem is more easily solved using the logarithm of $x_N(n_1, n_2, ..., n_s)$,

$$\log x_N(n_1, n_2, ..., n_s) = n_1 \log p_1 + n_2 \log p_2 + ... + n_s \log p_s. \qquad (5)$$

Each set of values $(n_1, n_2, ..., n_s)$ is a point with integer coordinates in a $s-$dimensional space. Since $n_1 + n_2 + ... + n_s = N$, all the points are in a subspace of dimension $s - 1$. The problem of the rank is reduced to find a path between the maximal rank point (with coordinates $(N, 0, 0, ..., 0)$) to the minimum $(0, 0, 0, ..., N)$ in such a way that $\log x_N(n_1, n_2, ..., n_s)$ decreases in

Fig. 6. Path of decreasing rank in the n_1, n_2 and n_3 space, for $N = 15$ and three random numbers $p_1 = 0.5202, p_2 = 0.3125$, $p_3 = 0.1673$.

each step. For $s = 2$, the solution is easy to find. Using that $n_1 + n_2 = N$,

$$x_N(n_1, n_2) = x_N(n_2) = p_1^{N-n_2} p_2^{n_2}, \tag{6}$$

from where it follows that the range is given by $r = n_2 + 1$. Then,

$$x_N(r) = p_1^N \left(\frac{p_2}{p_1}\right)^{r-1} = p_1^N e^{-A(r-1)}, \tag{7}$$

with $A = |\ln(p_2/p_1)|$. Eq. (7) shows that the numbers decrease in an exponential way as a function of the rank.

The case $s = 3$ can be easily visualized in Fig. 6, where the points in the integer lattice defined by Eq. (5) are shown as circles.

A path between points of decreasing $\log x_N(n_1, n_2, ..., n_s)$ is indicated as a line that joins the lattice points in Fig. 6, for a given set of numbers p_1, p_2 and p_3. Figure 7 shows how the values of n_1, n_2 and n_3 vary as a function of the range. A very complicated oscillatory pattern is seen , although a well defined envelope is also observed. This envelope is in fact the key to solve the problem, since it is the responsible of the ranking behavior. Notice also that all paths always start at $(N, 0, 0)$ and finish at $(0, 0, N)$, since $\log p_1 > \log p_2 > \log p_3$.

In general, since the index n_j is a function of the rank r, we can write that $n_j = n_j(r)$ where r is just the number of steps used to go from the point $(N, 0, ..., 0)$ to a certain $(n_1, n_2, n_3, ..., n_s)$. It follows that,

$$\log x_N(r) = n_1(r) \log p_1 + n_2(r) \log p_2 + ... + n_s(r) \log p_s \tag{8}$$

Fig. 7. Values of n_1 (thin solid line), n_2 (grey line) and n_3 (solid bold line) as a function of the rank, for $N = 20$ and $p_1 = 0.5202, p_2 = 0.3125$, $p_3 = 0.1673$.

Fig. 8. Path of decreasing ranks in the n_2 and n_3 plane for $p_1 \sim p_2 \gg p_3$, where the n_1 coordinate was eliminated using that $n_1+n_2+n_3 = N$. The dotted line corresponds to all the $n_{2MAX}(r)$, which defines the envelope of the ranking sequence.

The task is reduced to find the functions $n_j(r)$ for a given set $\{p_j\}$. Consider again the case of an initial set of three numbers, $s = 3$. Using that $n_1+n_2+n_3 = N$, $\log x_N(r)$ can be written as,

$$\log x_N(r) = N \log p_1 + n_2(r) \log \delta_{21} + n_3(r) \log \delta_{31}. \tag{9}$$

with $\delta_{21} = p_2/p_1$ and $\delta_{31} = p_3/p_1$. The solution for any set p_1, p_2 , p_3 is complicated, because some paths are not periodic. However, one can work out first the cases $p_1 \sim p_2 \gg p_3$ and $p_1 \gg p_2 \sim p_3$ that give insights about how to treat others.

Let us first consider the limit $p_1 \sim p_2 \gg p_3$, and $\delta_{21}^2 \gg \delta_{31}$. The corresponding path is easy to find because it is similar to an odometer with an increased range after each turn, as seen in Fig. 8, due to the hierarchy $1 > \delta_{21} > \delta_{21}^2 > \delta_{31} > \delta_{21}\delta_{31} > \delta_{31}^2 > ... > \delta_{31}^N$. For example, when $N = 2$ this leads to the following table that contains the number $x_N(r)$ as a function of the rank, and the corresponding path given by n_2 and n_3,

$x_N(r)$	n_2	n_3	r	$n_{2M}(r)$
p_1^2	0	0	1	–
$p_1^2\delta_{21}$	1	0	2	–
$p_1^2\delta_{21}^2$	2	0	3	2
$p_1^2\delta_{31}$	0	1	4	–
$p_1^2\delta_{21}\delta_{31}$	1	1	5	1
$p_1^2\delta_{31}^2$	0	2	6	0

The sequence of the path goes as follows, first $n_2(r)$ is increased one by one as n_3 remains constant, until it reaches a maximal value called $n_{2MAX}(r)$ which in fact determines the envelope of the ranking sequence and thus the basic shape of the curve $x_N(r)$ (the envelope that contains $n_{2MAX}(r)$ is shown in Fig. 8 as a dotted line). Once $n_2(r)$ increases from zero to $n_{2MAX}(r)$, a new cycle begins with $n_2(r) = 0$ and $n_3(r+1) = n_3(r)+1$. As a result, the number of steps r to reach $n_{2MAX}(r)$ is given by,

$$R - r \approx n_{2MAX}(r) + \sum_{j=1}^{n_{2MAX}(r)} j = n_{2MAX}(r) + \frac{n_{2MAX}(r)\,(n_{2MAX}(r)+1)}{2}, \quad (10)$$

where R is the maximal rank. Then,

$$n_{2MAX}(r) \approx N\left(1 - \frac{r}{R}\right)^{1/2} \quad (11)$$

The corresponding value of $n_3(r)$ can be obtained from the condition $n_2 + n_3 \leq N$. Finally, the number as a function of the rank is given by,

$$x_N(r) \approx \left[p_1 \left(\frac{p_2}{p_1}\right)^{\left(1-\frac{r}{R}\right)^{1/2}} \left(\frac{p_3}{p_1}\right)^{1-\left(1-\frac{r}{R}\right)^{1/2}}\right]^N. \quad (12)$$

Figure 9 shows the excellent agreement between Eq. (12) and the curve obtained for $p_1 = 0.5250$, $p_2 = 0.4250$, $p_3 = 0.000047$. Furthermore, Eq. (12) can

Fig. 9. Numerical results for the ranking of the succesive product of three numbers such that $p_1 \sim p_2 \gg p_3$. The smooth line is the prediction using Eq. (12).

be written as an stretched exponential as follows,

$$x_N(r) \approx p_3^N \exp\left[D\left(1-\frac{r}{R}\right)^{1/2}\right], \tag{13}$$

with $D = N\,|\log(p_2/p_3)|$ and R is the maximal value of r. Notice in Fig. 9 how this formula works better as the rank approaches R.

The case $p_1 \gg p_2 \sim p_3$ can be tackled in a similar way. The result is,

$$x_N(r) \approx p_1^N \exp\left[-E\left(\frac{r}{R}\right)^{1/2}\right]. \tag{14}$$

with $E = N\,(\log(p_1/p_3) - \log(p_2/p_3))$, and as shown in Fig. 10, the agreement is also good, specially for low values of r.

Now consider the general case in which p_1,p_2 and p_3 have the same order of magnitude, as in Fig. 5, where two tails appears, one for small r and the other at r near R. The tail at low r is produced basically by the hierarchy in the biggest probabilities, *i.e.*, by numbers where $n_1 \sim N$. In a similar way, the tail for r near R is produced by the lowest probability hierarchy, $n_3 \sim N$. These dominant factors are due to large statistical deviations and are the origin of the long tails in the otherwise power law observed in the ranks. The main effect in these tails when $p \approx p_2 \approx p_3$ is that the sequence of ordering is not uniform as can be observed in Fig. 5, for which a very complicate path appears. As a result, Eq. (10) changes with the apparition of new subcycles in the rank path. These changes are the result of the increasing number of cycles in the odometer that we have discussed, as is also clear from the change in the exponents that are transformed from 1 to 1/2 as s goes from $s = 2$ to $s = 3$.

Fig. 10. Ranking of the succesive product of three numbers such that $p_1 \gg p_2 \sim p_3$, for $p_1 = 0.99999$, $p_2 = 6.2 \times 10^{-6}$, $p_3 = 3.8 \times 10^{-6}$. The dashed line is the prediction made from Eq. (14), compared with the numerical result for $N = 100$ iterations (solid line).

Eq. (13) is thus transformed into a generalized expression,

$$x_N(r) \approx p_3^N \exp\left[D\left(1 - \frac{r}{R}\right)^{\beta}\right] \tag{15}$$

in which β is a yet unknown exponent, always less than one. In a similar way, Eq. (14) should be replaced by,

$$x_N(r) \approx p_1^N \exp\left[-E\left(\frac{r}{R}\right)^{\alpha}\right]. \tag{16}$$

with $\alpha < 1$. These generic exponents for the tails also appear for $s > 3$ since from the polynomial equivalent to Eq. (10), one gets $\alpha \approx \beta \approx 1/(s-1)$. A simple procedure to combine the tails represented by Eq. (15) and Eq. (16) is obtained by making the observation that for a given tail, only one stretched exponential produces a curved tails in a semi-log plot, while the other tends toward a constant, *i.e.*, if we consider the derivative of Eq. (15),

$$\left(\frac{d \ln x_N(r)}{dr}\right) = -\frac{\beta D}{R}\left(1 - \frac{r}{R}\right)^{\beta - 1} \tag{17}$$

is clear that $x'_N(r)$ is nearly a constant if $r \ll 1$, corresponding to the limit in which Eq. (16) has greater curvature. Analyzing the limit $r \to R$ gives a similar result,

$$\left(\frac{d \ln x_N(r)}{dr}\right) = -\frac{\alpha E}{R}\left(\frac{r}{R}\right)^{\alpha - 1}. \tag{18}$$

From these considerations, a simple way to produce a function with the required dependences when $r \to R$ and $r \to 1$ is the following,

Fig. 11. A plot of Eq. (19) using $C_1 = 2$, $D = 1$, $E = 1$, $\beta = 1$ and $\alpha = 1$.

$$x_N(r) \approx C_1 \exp\left[D\left(1 - \frac{r-1}{R}\right)^{\beta}\right] \exp\left[-E\left(\frac{r}{R}\right)^{\alpha}\right], \tag{19}$$

where C_1 is a constant. A plot of the previous expression is presented in Fig. 11, showing the basic shape of the studied beta function.

Finally, Eq. (19) can be simplified when many states are present since for $s \gg 1$, $\alpha \approx \beta \approx 1/(s-1)$ and thus $\alpha \to 0$ and $\beta \to 0$. Then, by using the observation about the derivatives that appears in Eq. (17) and Eq. (18), one can approximate the derivatives like in Eq. (17) as follows,

$$\left(\frac{d \log x_N(r)}{dr}\right) = -\frac{\beta D}{R}\left(1 - \frac{r}{R}\right)^{\beta-1} \approx -\frac{\beta D}{R}\left(1 - \frac{r}{R}\right)^{-1}. \tag{20}$$

A similar thing can be done in the tail $r \to 1$, for which α can be neglected with respect to one in Eq. (18). Combining both tails in a sole expression we get,

$$\left(\frac{d \log x_N(r)}{dr}\right) \approx -\frac{\beta D}{R}\left(1 - \frac{r}{R}\right)^{-1} - \frac{\alpha E}{R}\left(\frac{r}{R}\right)^{-1}.$$

By integrating the previous equation, we finally obtain the beta-like function given by Eq. (1), where the exponents a and b are given by,

$$a = \alpha E \quad \text{and} \quad b = \beta D \tag{21}$$

Thus, the beta-like function is obtained when we have a large number of states in the system. Notice how the parameters a and b are determined mainly by the behavior in the tails.

5 Conclusions

In conclusion, we found a simple formula that allows to fit many different rank phenomena. This formula shows that there is a certain universality at the tails, explained by considering the ranking of a multiplicative process. We have shown that such problem is equivalent to an algebraic problem: find the rank of the successive product of numbers. A task that remains is to how to get the coefficients a and b from physical principles, using for example master equations and the concept of multiscaling modelling. A key observation for such study is that for expansion-modification algorithms in DNA models, $a > b$ if the expansion probability of the genetic code is bigger the than mutation rate [21]. Thus, a and b represent the relative influence of two general mechanisms, where each of them dominate at a given tail. According to some preliminary results, a seems to be related with a certain funnel type of energy landscape, as in protein folding, which leads to a certain deterministic sequence, while b is associated with a many valley landscape, as seen in spin glasses. This last opposite effect provides much more variability in the sequence of results. Such correlation is consistent with associating b to the stochastic component of the dynamics and a with the most deterministic features [21]. In future work, we will elucidate with more detail such mechanisms.

Acknowledgments. This work was supported by DGAPA-UNAM project IN-117806, CONACyT 48783-F and 50368.

References

[1] Li W., Phys. Rev. E **43**, 5240 (1991), see also: Li W., http://www.nslij-genetics.org/wli/zipf/ (2003).

[2] M. Bretz, R. Zaretzki, S.B. Field, N. Mitarai and F. Nori, Europhysics Lett. **74** (2006) 1116 .

[3] G. Audi, O. Bersillon, J. Blachot and A. H. Wapstra, Nuclear Physics **A624** (1997) 124.

[4] S. Fortunato, A. Flammini, and F. Menczer, Phys. Rev. Lett. **96** (2006) 218701.

[5] A.C.C. Yang, S.S. Hseu, H.W Yien, A.L. Goldberger, and C.K. Peng, Phys. Rev. Lett. **90** (2003) 108103.

[6] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabasi, Nature **407** (2000) 651.

[7] Le Quan H., Sicilia-García E.I., Minj J. and Smith F.J., *Proceedings of the 17th. International Conference on Computer Lingustics,* Montreal, (2002).

[8] J. Laherrere and D. Sornette, Eur. Phys. J.B. **2** (1998) 525 .

[9] E.W. Montroll and M.F. Shlesinger, J. of Statistical Physics **32** (1983) 209.

[10] A.N. Kolmogorov, Dokl. Akad. Nauk SSSR **30** (1941) 299 (reprinted in Proc. R. Soc. Lond. A **434** (1991) 9).

[11] A.N. Kolmogorov, Dokl. Akad. Nauk SSSR **32** (1941) 16 (reprinted in Proc. R. Soc. Lond. A **434** (1991) 15).

[12] A. Kolmogorov, J. Fluid. Mech. **13** (1962) 82 .

[13] Z. Warhaft, Annu. Rev. Fluid Mech. **32**, (2000) 203.

[14] L.G. Moyano, C. Tsallis and M. Gell-Mann, Europhys. Lett. **72** (2006) 355.

[15] J. A. Marsh, M.A. Fuentes, L.G. Moyano and C. Tsallis, Physica A **372** (2006) 183 .

[16] Popescu I., Glottometrics **6** (2003) 83.

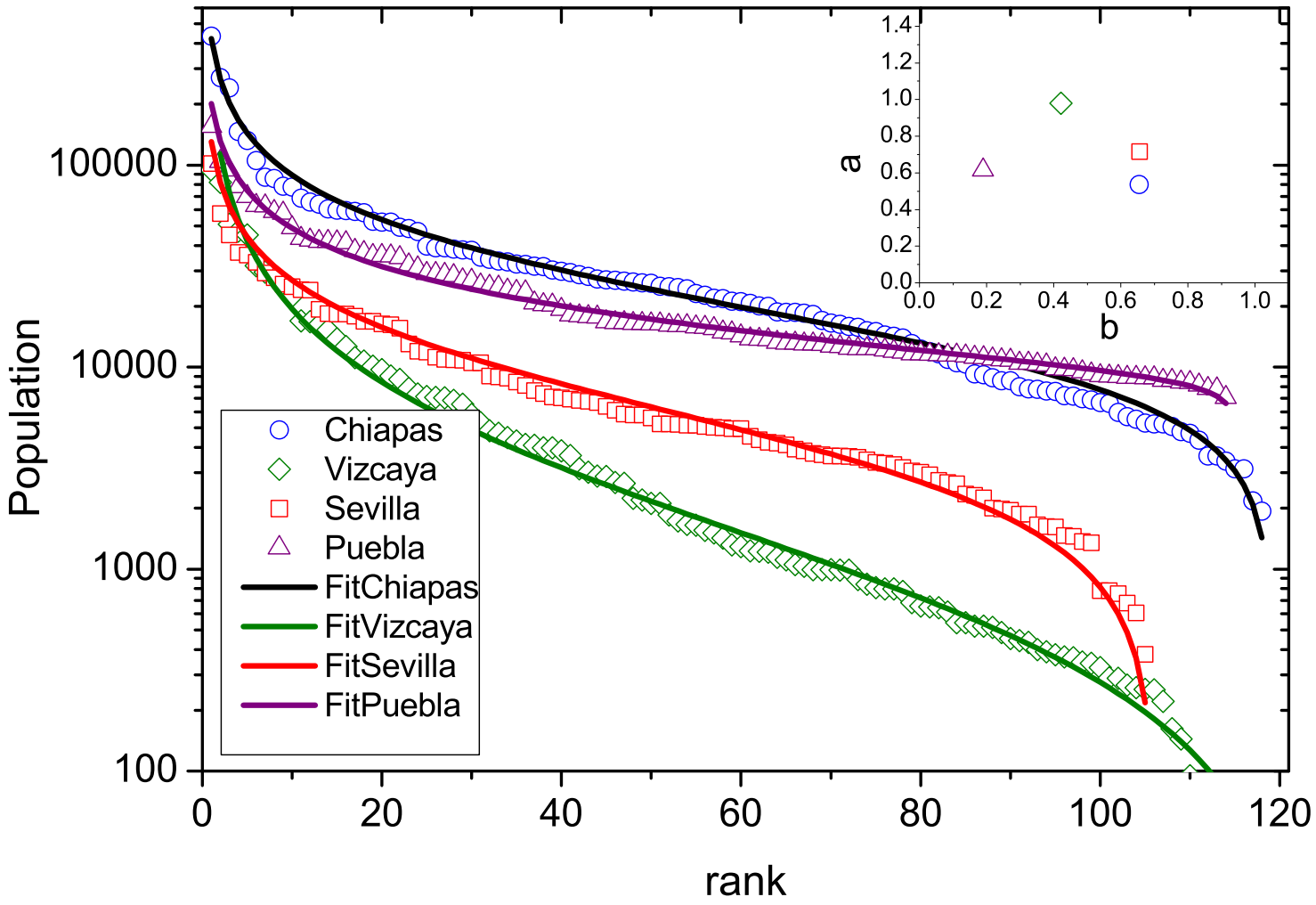
[17] Codon Usage Database, NCBI-GenBank http://www.kazuza.or.jp/codon

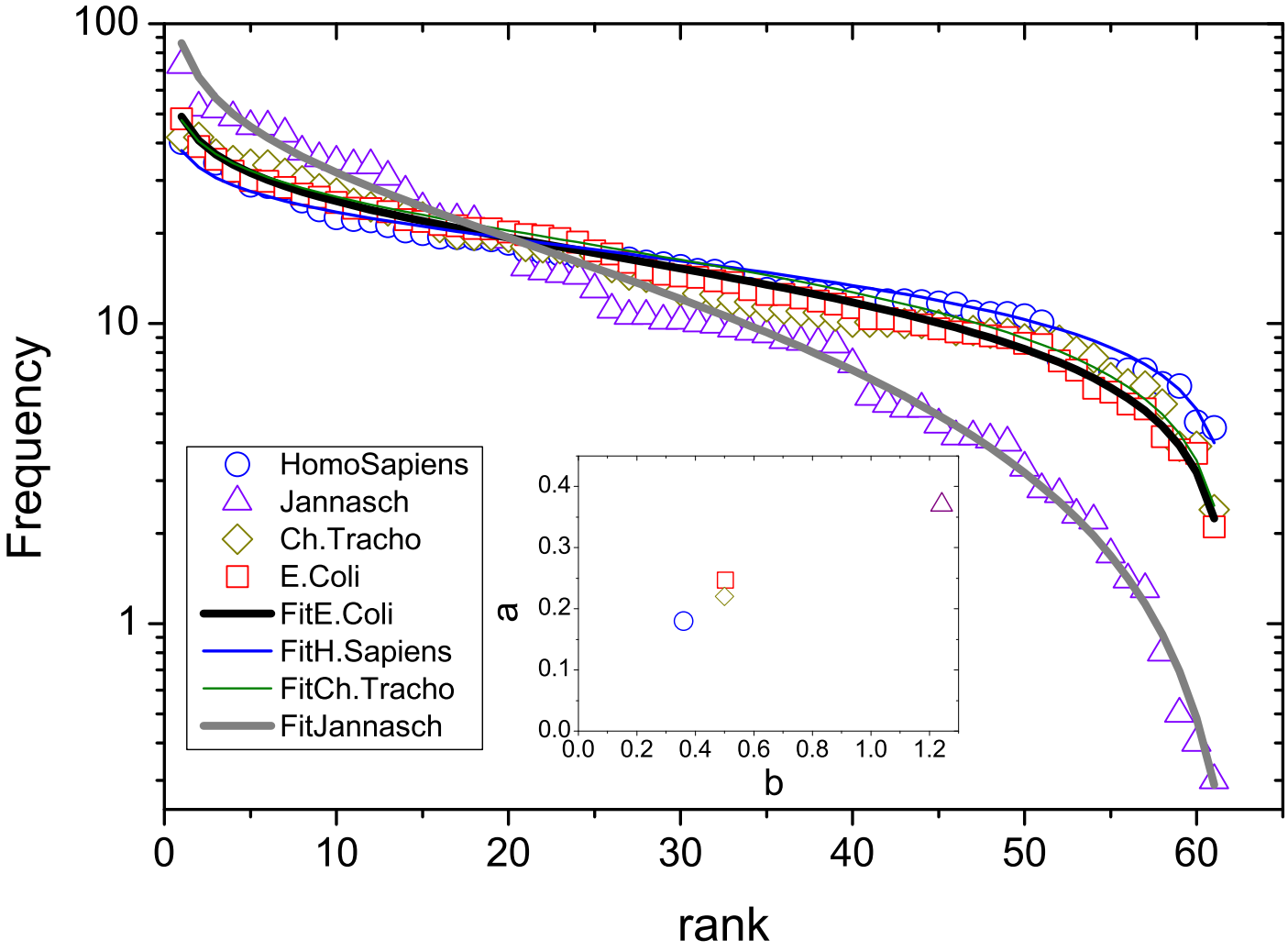
[18] G. Cocho, G. Martínez-Mekler, to be published.

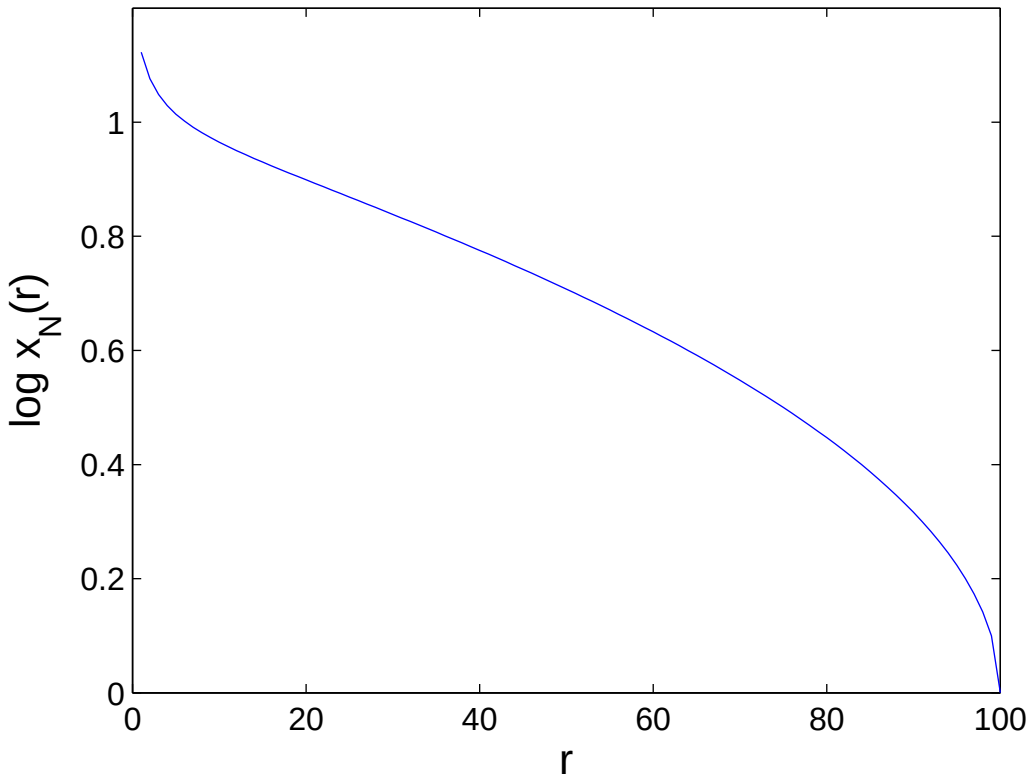
[19] S.C. Manrubia, D.H. Zanette, Phys. Rev. E **59** (1999) 4945 .

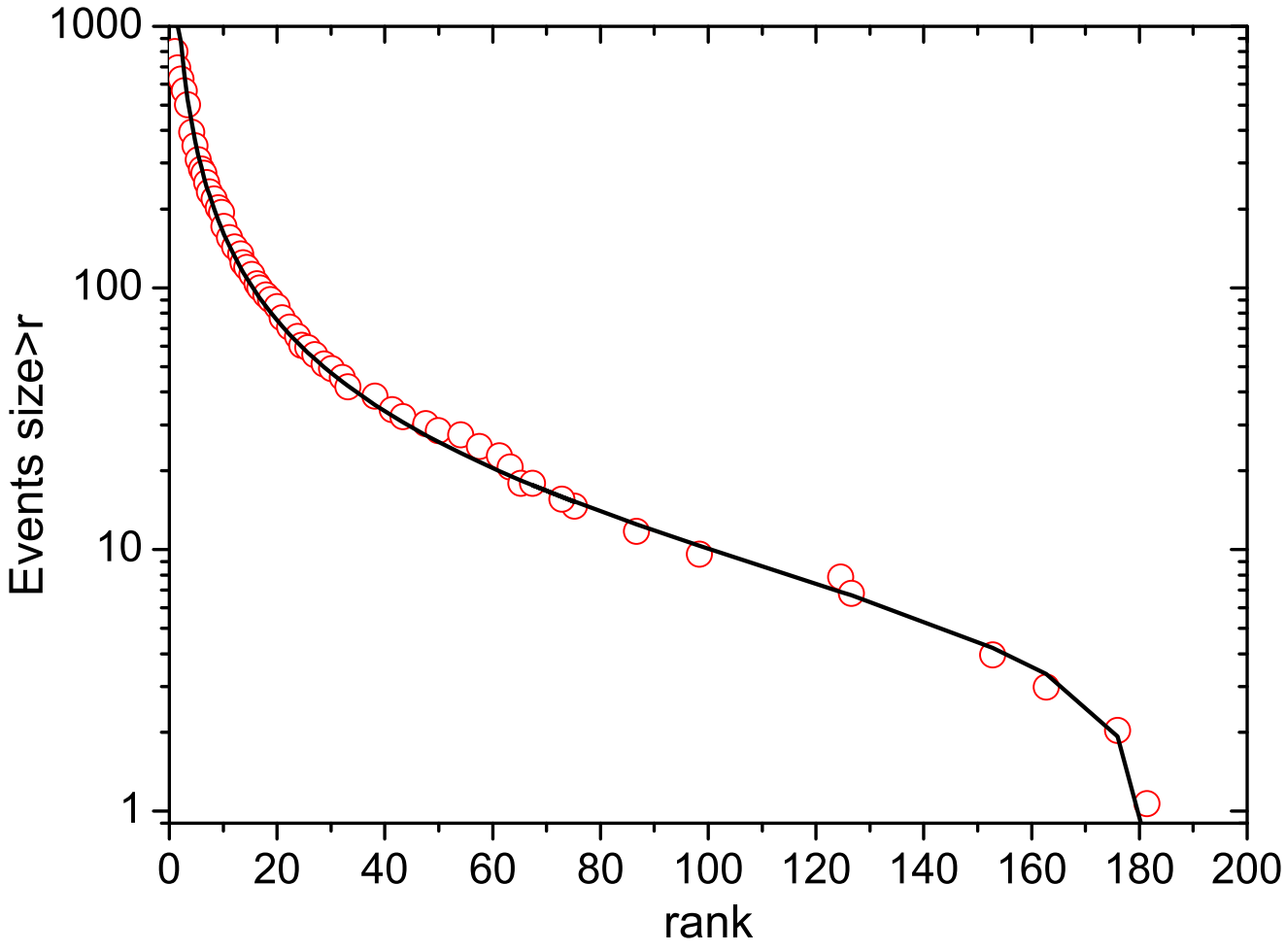
[20] G. Naumis, G. Cocho, to be published.

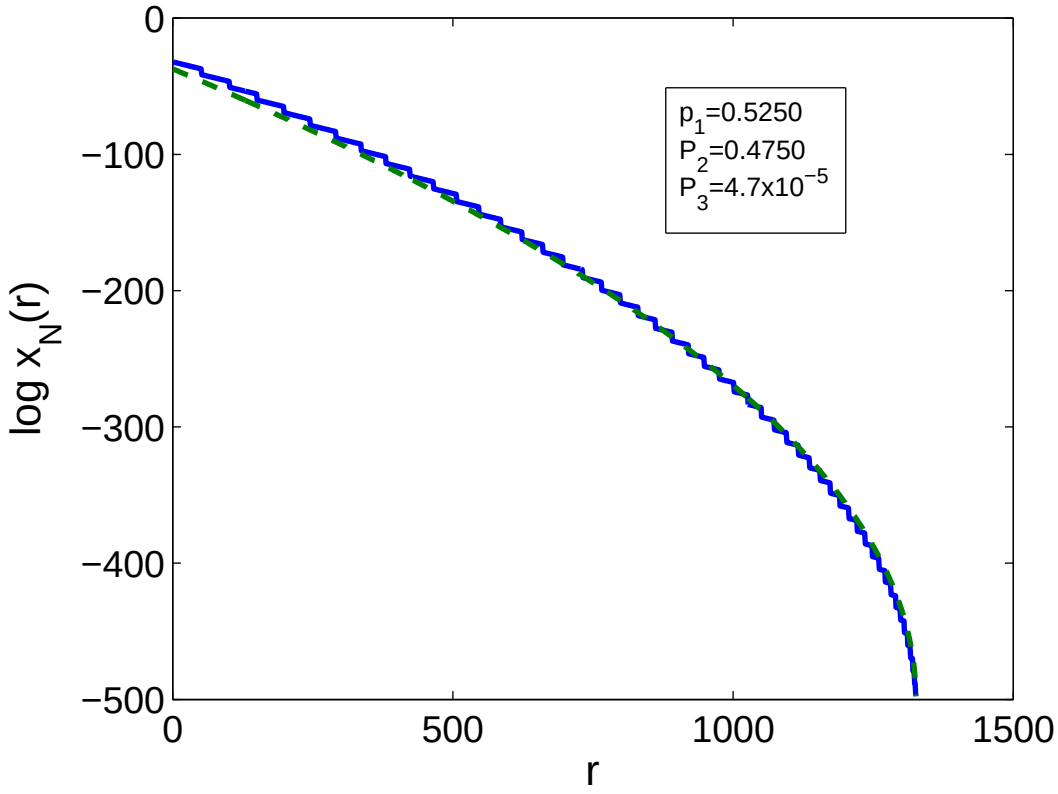
[21] Mansilla R, Cocho G., Complex Systems. **12** (2000) 207.

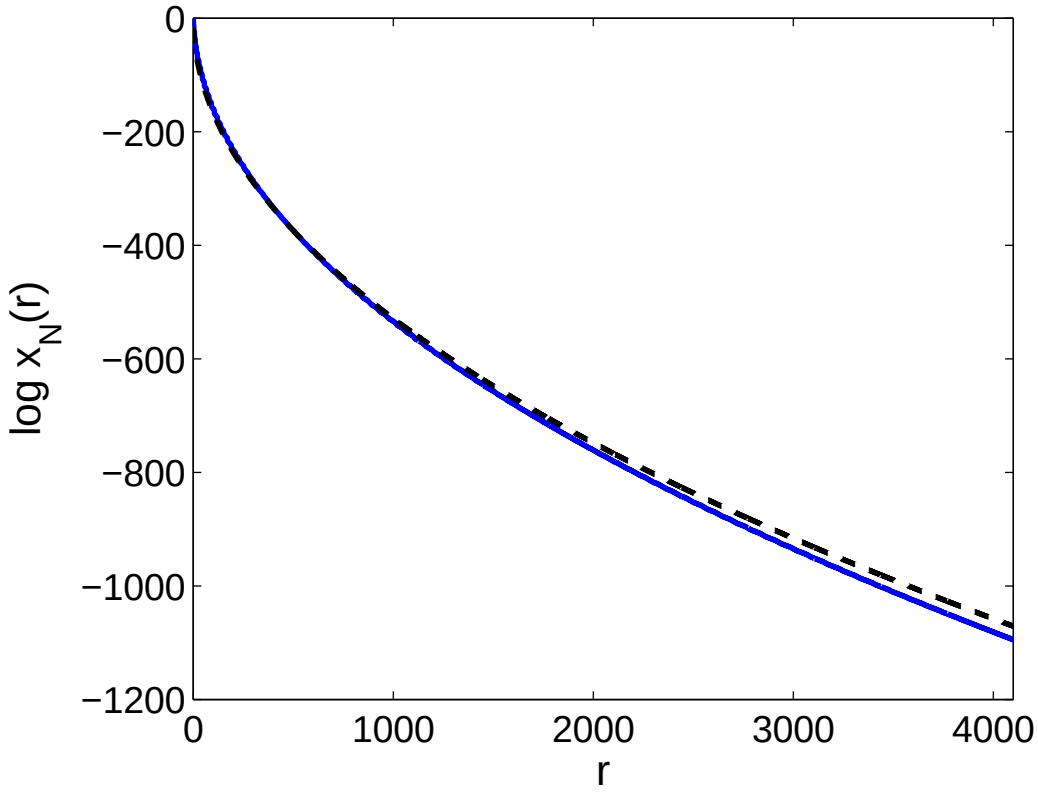

Population
rank
100000
10000
1000
100
0
20
40
60
80
100
120
Chiapas
Vizcaya
Sevilla
Puebla
FitChiapas
FitVizcaya
FitSevilla
FitPuebla
a
b
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

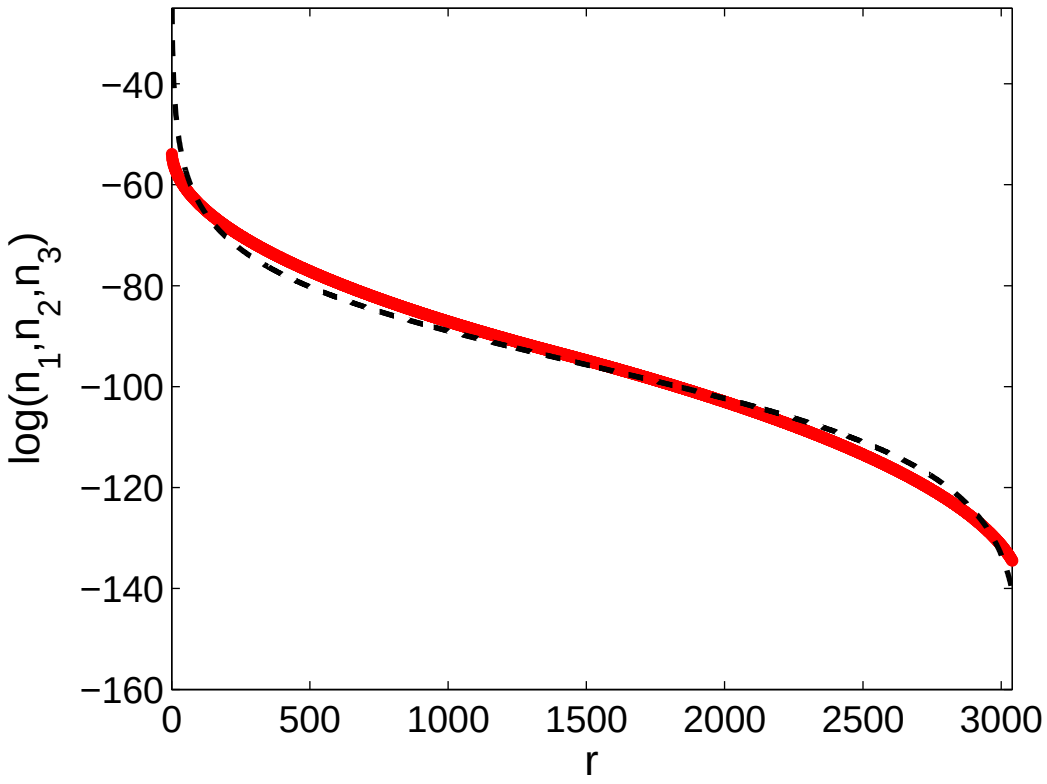

100
10
1
Frequency
0
10
20
30
40
50
60
rank
HomoSapiens
Jannasch
Ch.Tracho
E.Coli
FitE.Coli
FitH.Sapiens
FitCh.Tracho
FitJannasch
a
b
0.0
0.1
0.2
0.3
0.4
0.0
0.2
0.4
0.6
0.8
1.0
1.2

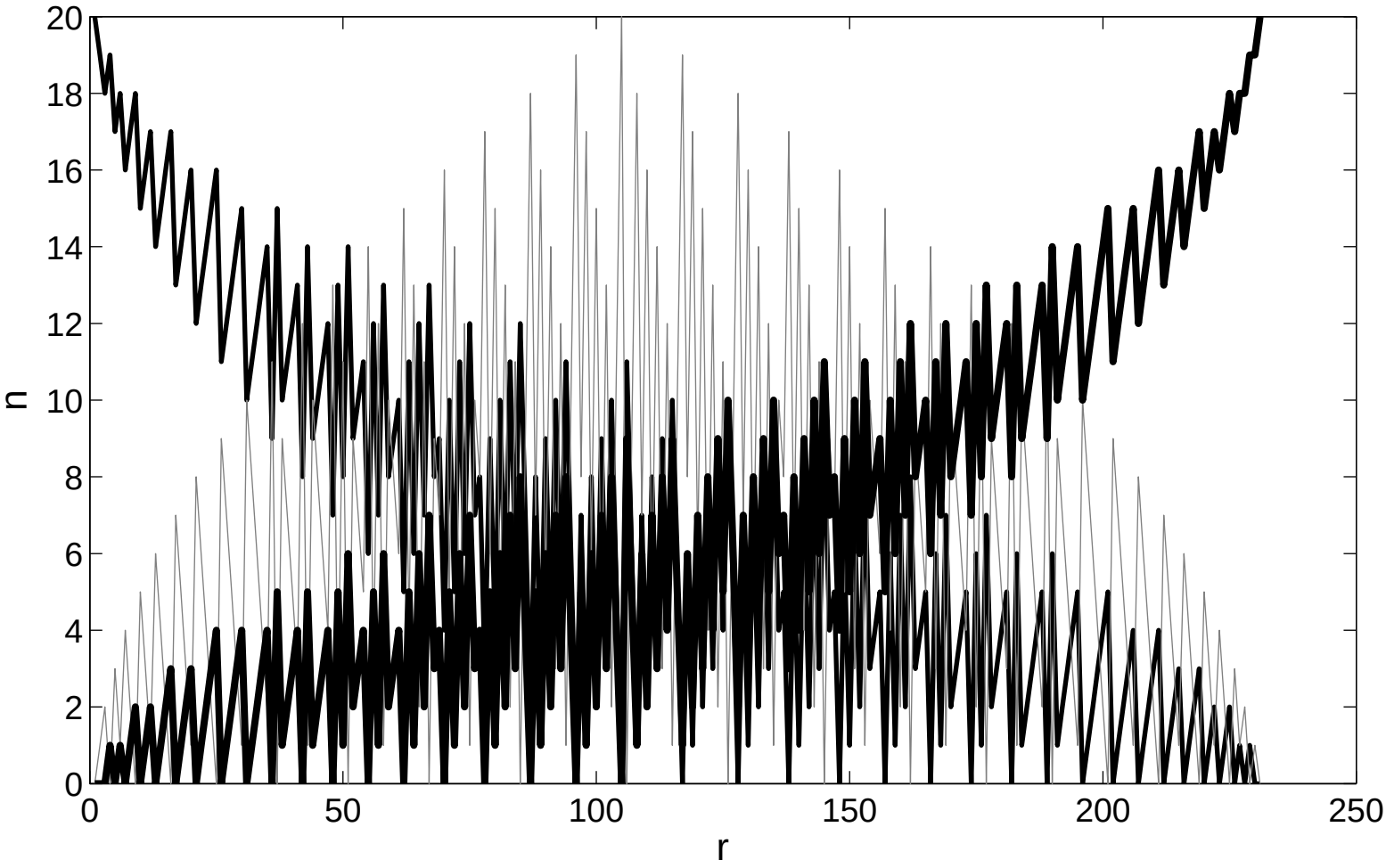

log $x_N(r)$
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
r

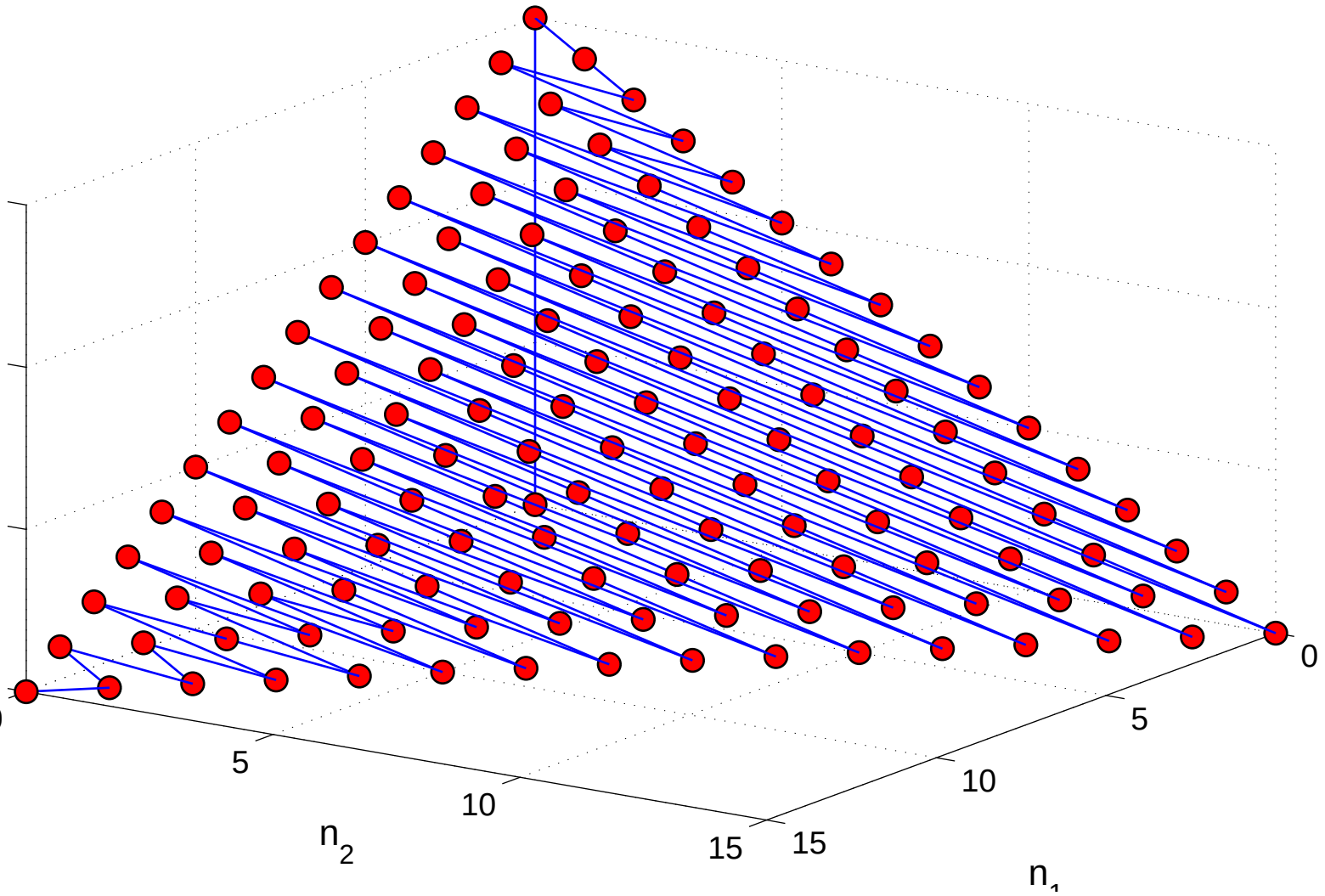

1000
100
10
1
Events size>r
0
20
40
60
80
100
120
140
160
180
200
rank

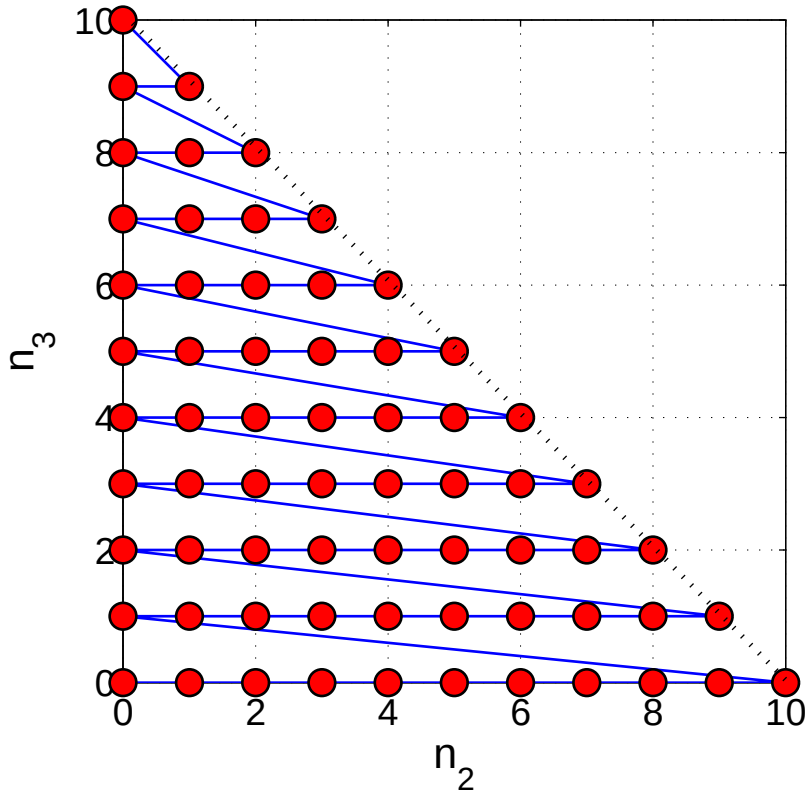

0
−100
−200
−300
−400
−500
0
500
1000
1500
r
log x_N(r)
p_1=0.5250
P_2=0.4750
P_3=4.7x10^−5


0
−200
−400
−600
−800
−1000
−1200
log x_N(r)
0
1000
2000
3000
4000
r


−40
−60
−80
−100
−120
−140
−160
log(n1,n2,n3)
0
500
1000
1500
2000
2500
3000
r


20
18
16
14
12
10
8
6
4
2
0
n
0
50
100
150
200
250
r


5
10
15
n_2
15
10
5
0
n_1


10
8
6
4
2
0
n3
0
2
4
6
8
10
n2